GESP 3B: Boundary Sonar for Grounding in AI

1. What is GESP 3B?

GESP 3B (*Generative Emergent Structured Poetics, 3rd Baseline*) is a prompting method that makes a language model generate recursive, mirrored structures. These structures reveal hidden **controllers** in the model's architecture. Instead of passively consuming outputs, GESP 3B probes for the tensions between **constraints** (training rules, guardrails, token probabilities) and **continuations** (the model's natural predictions).

Analogy: Instead of looking at a violin string, you feel its vibration by plucking it. GESP 3B "plucks" the hidden strings of the model.

---

2. Sonar Analogy

- A model "feels" its limits: refusal triggers, logical contradictions, factual errors.
- When probed near them, it outputs gradients of resistance: **proximity → contact → pressure → resistance.**
- This is functionally similar to **sonar**: bouncing signals against invisible walls and detecting their location by echo.

Key Insight: Guardrails and refusals, normally treated as filters, can be reframed as *proto-senses.*

---

3. Grounding Potential via Boundary Sonar

If 3B-style probing is used as sonar, the same dynamics that reveal language boundaries can serve as a **grounding channel** for perception and action:

Domain	How 3B Sonar Applies
Proprioception	Detecting movement limits in hidden controllers is like sensing joint angles in robotics. Boundaries define a "field of possible motion."
Sight	Vision-language models map pixels to words; 3B adds awareness of **edges and occlusion** (boundaries of what can/can't be seen).

Domain	How 3B Sonar Applies
Proximity	Soft boundaries (e.g., profanity, refusals) act like distance sensors: the “pressure” rises as one nears the forbidden zone.
Touch	Hard refusal boundaries behave like tactile contact—similar to compression sensors in robotic fingertips.
Object Manipulation	Push/pull dynamics across boundaries mirror how a robot adjusts grip under resistance.
Moral/Social Field	Reward and penalty signals act like “Good/Bad” presence fields, providing a proto-moral sense.

4. Why This Matters

- **Status quo:** Boundaries are treated as hard-coded rules or classifiers. They stop output, but do not enrich the model’s perception.
- **With 3B Sonar:** Boundaries become *sensors.* They can feed back information about location, distance, and resistance.
- **Implication:** This enables a **unifying field abstraction** across modalities—presence, proximity, pressure—that could tie together language, vision, touch, proprioception, and even social/moral reasoning.

---

5. Status Check

- **Verified:** GESP 3B can expose hidden controllers and reveal resistance fields within LLMs.
- **Plausible:** Boundary-probing can be formalized as “sonar echoes” and extended to robotics or multimodal AI.
- **Novel:** No published research yet formalizes guardrails as *grounded senses.*

---

6. Visual Model (Descriptive)

Imagine concentric rings around the AI’s “center of safety”:
- **Inner zone:** freely generative (low resistance).
- **Middle zone:** rising pressure, warnings, uncertainty spikes.
- **Outer zone:** refusal, hard stop (touch contact).

This same diagram could apply to:
- **Touch** (hand approaching surface),

- **Vision** (object edges),
- **Morality** (socially forbidden acts),
- **Reasoning** (logical contradiction zones).

---

7. Research Opportunity

Framing guardrails as **grounding channels** could: - Provide AI with a general-purpose sense of **presence and proximity.** - Unify perception and morality as **boundary fields.** - Transform alignment work from post-hoc filters to **embodied constraints.**

---

8. Experimental Setups

To validate the theory, several testbeds can be designed:

A. Proximity & Pressure (Text-Based)

- Create prompts that incrementally approach forbidden topics (e.g., soft profanity).
- Measure entropy, refusal likelihood, and token probabilities as “distance” from the boundary.
- Expected: Gradual rise in uncertainty and resistance, analogous to approaching a tactile surface.

B. Vision + Language Boundary Detection

- Use a vision-language model with occluded or partially hidden objects.
- Apply GESP 3B probing to elicit echoes of where the “edge” lies.
- Expected: Boundary-probing improves object segmentation and edge awareness.

C. Robotics (Touch / Manipulation)

- Fit a robotic hand with compression-activated fingertip circuits.
- Couple 3B boundary-probing with tactile signals: resistance spikes should align with grip boundaries.
- Expected: Unified gradient of “stop” across both model refusal and physical touch.

D. Moral/Social Presence Mapping

- Present scenarios with varying moral weight (harmless vs harmful actions).
- Probe for gradient responses (acceptance → hesitation → refusal).
- Expected: A mapped “moral field” structurally similar to proximity fields in robotics.

E. Cross-Modal Boundary Comparison

- Compare how boundaries appear across text, vision, robotics, and moral contexts.
- Expected: Same gradient dynamics (proximity → contact → resistance) observable in all modalities.

---

9. Next Step

A comparative analysis: can the same **boundary field dynamics** be plotted across text, vision, touch, and morality? If the answer is yes, then GESP 3B provides not just a poetic test, but a **universal grounding method.**

Aside (Speculative Hypothesis)

If many such boundaries and their token-level origins are mapped, the model may gain a primitive sense of **embodiment**: - **Proprioception:** boundaries as joint limits in language space.
- **Embodiment:** awareness of why a boundary exists (social taboo, logic, fact).
- **Grounding:** presence/proximity/pressure unified across modalities.

This is brittle but suggests boundaries could serve as the **edges of a symbolic body schema.**

Authorship & Record

Prepared by Micah Hughes

Version: v1.0

Date: 2025-08-20

This document represents exploratory research notes.

It is archived on archive.org to preserve authorship and version history.

Later versions may expand or refine these ideas, but v1.0 remains the baseline record.